Université Paris I, Panthéon - Sorbonne

Première Année Master M.A.E.F. 2015 – 2016

Econométrie II

Examen final, mai 2016

Examen de 3h00. Tout document ou calculatrice est interdit.

1. **(16 points)** On suppose que pour n et p deux entiers tels que $n \geq p+2$, on observe $Y = {}^t(Y_i)_{1\leq i\leq n}$ défini par:

$$Y_i = \theta_0 + \theta_1 X_i^{(1)} + \cdots + \theta_p X_i^{(p)} + \varepsilon_i \qquad \text{pour tout } i = 1, \cdots, n, \tag{1}$$

avec une famille connue de réels $(X_i^{(j)})_{1\leq i\leq n,\, 1\leq j\leq p}$, et telle que $X = \begin{pmatrix} 1 & X_1^{(1)} & \cdots & X_1^{(p)} \\ \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & \cdots & X_n^{(p)} \end{pmatrix}$ soit une matrice de rang $p+1$, $\theta = {}^t(\theta_j)_{0\leq j\leq p}$ un vecteur de nombres réels inconnus et $\varepsilon = {}^t(\varepsilon_i)_{1\leq i\leq n}$ un vecteur d'erreur de loi $\mathcal{N}\big(0\,,\, \sigma^2\, I_n\big)$ non observé, où $\sigma^2 > 0$ est inconnu et I_n est la matrice identité de taille n.
Pour M une matrice de taille $(n, p+1)$, on note $[M] = \{M\,\alpha,\ \alpha \in \mathbf{R}^{p+1}\}$ et P_A la matrice de projection orthogonale sur un sous-espace A de $\mathbf{R}^n$.

   (a) Pour A et B deux sous-espaces vectoriels de $\mathbf{R}^n$, montrer que $P_A\, P_B = P_B\, P_A = P_{A\cap B}$.

   (b) Ecrire le modèle sous une forme matricielle, et rappeler l'expression du vecteur de prédiction $\widehat{Y} = (\widehat{Y}_i)_{1\leq i\leq n} = X\,\widehat{\theta}$ obtenu par moindres carrés ordinaires. Démontrer que $\widehat{Y}$ a pour loi $\mathcal{N}_n\big(0\,,\, \sigma^2\, X({}^t X\, X)^{-1\,t} X\big)$.

   (c) Rappeler en la justifiant la loi de $\widehat{\sigma}^2$ estimateur non biaisé de σ^2.

   (d) On veut tester si l'observation Y_n n'est pas une donnée aberrante. On considère ainsi le problème de test H_0: ε_n suit la loi $\mathcal{N}(0\,,\, \sigma^2)$, contre H_1: ε_n ne suit pas la loi $\mathcal{N}(0\,,\, \sigma^2)$. Soit $X_{(n)}$ la matrice X où l'on remplace la dernière ligne par des 0 et $E_n = \mathrm{Vect}\big({}^t(0, \cdots, 0, 1)\big)$. Montrer que $E_n \subset [X_{(n)}]^{\perp}$. Montrer que $\widehat{\sigma}^2_{(n)}$ l'estimateur non biaisé de σ^2 à partir de l'échantillon privé de l'individu n s'écrit $\frac{1}{n-p-2}\, \|P_{E_n^{\perp}}\, P_{[X_{(n)}]^{\perp}}\varepsilon\|^2$.

   (e) Montrer que $\widehat{\varepsilon}_n = Y_n - \widehat{Y}_n = (0, \cdots, 0, 1)\, P_{E_n}\, P_{[X]^{\perp}}\varepsilon$. En déduire que $\varepsilon'_n = \dfrac{Y_n - \widehat{Y}_n}{\big(\widehat{\sigma}^2_{(n)}\,(1 - p_{nn})\big)^{1/2}}$ suit une loi de student à $n-p-2$ degrés de liberté, où $P_{[X]} = (p_{ij})_{1\leq i,j\leq n}$. Comment utiliser ε'_n pour accepter ou non H_0?

   (f) On suppose que l'on observe $X_{n+1} = \big(1, X_{n+1}^{(1)}, \ldots, X_{n+1}^{(p)}\big)$, vecteur de réels connus, et l'on veut prédire Y_{n+1}. Montrer qu'une prédiction de Y_{n+1} est $\widetilde{Y}_{n+1} = X_{n+1}({}^t X\, X)^{-1\,t} X\, Y$ et montrer que $\mathbb{E}(\widetilde{Y}_{n+1}) = \mathbb{E}(Y_{n+1})$. Déterminer la loi de $\widetilde{Y}_{n+1}$, puis montrer que $\dfrac{\widetilde{Y}_{n+1} - \mathbb{E}(Y_{n+1})}{\big(\widehat{\sigma}^2\, X_{n+1}({}^t X\, X)^{-1\,t} X_{n+1}\big)^{1/2}}$ suit une loi que l'on précisera. En déduire un intervalle de prédiction à 95% pour Y_{n+1}. Sans l'hypothèse gaussienne sur ε, quel intervalle aurait-on pu obtenir et sous quelles hypothèses?

   (g) On se place dans le simple cadre où $p = 1$, et sans l'hypothèse de gaussianité de ε. Ecrire $\widehat{\theta}$ en fonction des $x_i = X_i^{(1)}$ et des Y_i (on utilisera les notations $\overline{x}_n = \frac{1}{n}\,(x_1 + \cdots + x_n)$ et $\overline{Y}_n = \frac{1}{n}\,(Y_1 + \cdots + Y_n)$). Donner une hypothèse suffisante pour que $\widehat{\theta}$ satisfasse un théorème de la limite centrale que l'on précisera. Donner alors explicitement la statistique du test de Fisher global du modèle ainsi que sa limite. Donner ensuite l'intervalle asymptotique de prédiction à 95% pour Y_{n+1} en fonction de x_{n+1}. Que se passe-t-il quand $x_{n+1} \to \infty$?

Proof. (a)

□

2. **(10 points)** Exercice de TP utilisant le logiciel `R`

On considère un jeu de données réelles mettant en relation une mesure quantitative du degré d'avancement du diabete (variable `prog`) et 10 variables cliniques dans une cohorte de 442 patients. Chacun des 442 patients est décrit par les variables suivantes :

- `prog`: l'indice de progression de la maladie (le plus grand, le plus en progrès)
- `age`: l'âge du patient
- `sex`: le sexe du patient, codé numériquement (1, homme ou 2, femme)
- `bmi`: l'indice de masse corporelle
- `map`: la pression artérielle moyenne
- `ser1` - `ser6`: 6 mesures sérologiques

On a séparé au préalable et aléatoirement le jeu de données en deux sous-ensembles: `diabetes1` et `diabetes2`.

(a) On cherche à construire un modèle linéaire avec de bonnes propriétés prédictives pour la variable `prog`. On a ainsi tapé les commandes suivantes avec le logiciel R:

```
Diab1=read.table("C:/Donnees/diabetes1.txt",header=TRUE)
Diab1$sex=as.factor(Diab1$sex)
attach(Diab1)
reg1=lm(prog~age+bmi+map+ser1+ser2+ser3+ser4+ser5+ser6+sex)
summary(reg1)
```

Voici le début de `Diab1`:

```
> Diab1
    prog age sex  bmi    map ser1  ser2 ser3 ser4   ser5 ser6
1    151  59   2 32.1 101.00  157  93.2 38.0 4.00 4.8598   87
3    141  72   2 30.5  93.00  156  93.6 41.0 4.00 4.6728   85
6     97  23   1 22.6  89.00  139  64.8 61.0 2.00 4.1897   68
9    110  60   2 32.1  83.00  179 119.4 42.0 4.00 4.4773   94
10   310  29   1 30.0  85.00  180  93.4 43.0 4.00 5.3845   88
:     :   :   :  :     :      :    :    :    :     :      :
```

Une partie des résultats obtenus est présente ci-dessous:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -319.8486    84.3700  -3.791 0.000202 ***
age            0.1468     0.3097   0.474 0.635981
bmi            6.2159     1.0607   5.860 2.02e-08 ***
map            1.1904     0.3172   3.753 0.000232 ***
ser1          -0.3994     0.7728  -0.517 0.605918
ser2          -0.2271     0.7468  -0.304 0.761353
ser3           0.2344     0.9817   0.239 0.811551
ser4          20.6894     8.1506   2.538 0.011943 *
ser5          41.5733    21.7245   1.914 0.057174 .
ser6           0.1199     0.3858   0.311 0.756264
sex2         -22.9688     8.2902  -2.771 0.006154 **

Residual standard error: 52.35 on 189 degrees of freedom
Multiple R-squared:  0.5669,    Adjusted R-squared:  0.544
F-statistic: 24.74 on 10 and 189 DF,  p-value: < 2.2e-16
```

Questions I.1: Pourquoi a-t-on tapé la deuxième commande? Montrer mathématiquement que dans le cadre de cette variable à deux modalités il n'était pas obligatoire de la taper. Combien y-a-t-il d'invidus dans cette base de données? Ecrire précisément le modèle sous forme matricielle en précisant exactement la première ligne de la matrice X. En notant les paramètres θ_j, $j = 0, \ldots, p$, dans l'ordre donné ci-dessus, que vaut numériquement $\widehat{\theta}_4$? Ecrire mathématiquement ce qu'est le nombre 0.3097. *De ces résultats, pourriez-vous conclure statistiquement que plus on est en surpoids, plus la maladie progresse? Que pensez-vous de cette régression?*

(b) On tape ensuite les commandes:

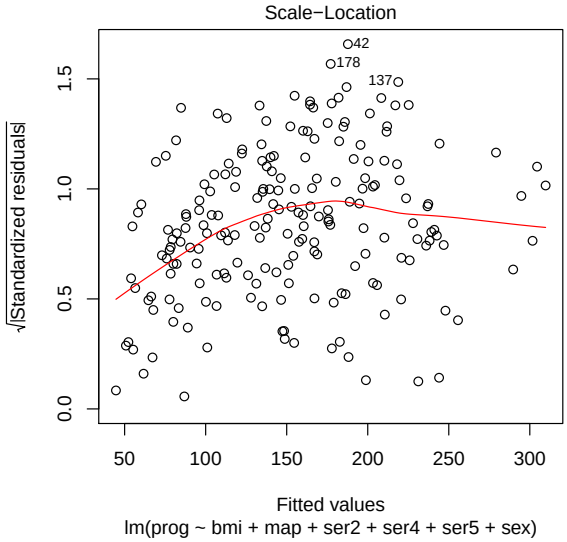
```
library(leaps)
Z=matrix(c(age,bmi,map,ser1,ser2,ser3,ser4,ser5,ser6,sex),ncol=10);
colnames(Z)=c("age","bmi","map","ser1","ser2","ser3","ser4","ser5","ser6","sex");
r=leaps(Z,prog)
t=(r$Cp==min(r$Cp))
colnames(Z)[r$whi[t]]
reg2=lm(prog
```

Voici les résultats obtenus:

```
> colnames(Z)[r$whi[t]]
[1] "bmi"  "map"  "ser2" "ser4" "ser5" "sex"

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -291.5707    39.0335  -7.470 2.71e-12 ***
bmi            6.4745     0.9778   6.621 3.44e-10 ***
map            1.2302     0.3008   4.089 6.35e-05 ***
ser2          -0.6142     0.1653  -3.716 0.000265 ***
ser4          22.0946     4.6624   4.739 4.16e-06 ***
ser5          31.5945     9.1086   3.469 0.000645 ***
sex2         -21.9458     8.1080  -2.707 0.007405 **

Residual standard error: 51.92 on 193 degrees of freedom
Multiple R-squared:  0.5651,    Adjusted R-squared:  0.5516
F-statistic:  41.8 on 6 and 193 DF,  p-value: < 2.2e-16
```



Questions I.2: Qu'a-t-on fait en tapant ces commandes? Que concluez vous quant aux résultats et graphe obtenus?

On tape maintenant les commandes suivantes:

```
library(MASS)
X=as.data.frame(Z)
reg3=stepAIC(lm(prog~.,data=X),k=log(200))
summary(reg3); plot(reg3)
```

Voici ce que l'on obtient à la fin des résultats:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -311.7870    36.4236  -8.560 3.19e-15 ***
bmi            8.1540     0.9918   8.221 2.67e-14 ***
ser5          53.7323     8.1519   6.591 3.90e-10 ***

Residual standard error: 55.9 on 197 degrees of freedom
Multiple R-squared:  0.4854,    Adjusted R-squared:  0.4801
F-statistic:  92.9 on 2 and 197 DF,  p-value: < 2.2e-16
```

Questions I.3: Qu'a-t-on fait en tapant ces commandes? Que concluez vous quant aux résultats obtenus?

On effectue maintenant ceci:

```
Diab2=read.table("C:/Donnees/diabetes2.txt",header=TRUE)
Diab2$sex=as.factor(Diab2$sex)
new=data.frame(age=Diab2$age,bmi=Diab2$bmi,map=Diab2$map,ser1=Diab2$ser1,ser2=Diab2$ser2,
```

```
ser3=Diab2$ser3,ser4=Diab2$ser4,ser5=Diab2$ser5,ser6=Diab2$ser6,sex=Diab2$sex)
pred1=predict(reg1,new); pred2=predict(reg2,new); pred3=predict(reg3,new)
MSE1=sqrt(mean((Diab2$prog-pred1)^2)); MSE2=sqrt(mean((Diab2$prog-pred2)^2));
MSE3=sqrt(mean((Diab2$prog-pred3)^2)); MSE1; MSE2; MSE3
```

Et on obtient:

```
> MSE1; MSE2; MSE3
[1] 56.98467
[1] 57.47944
[1] 57.73582
```

Questions I.4: Qu'a-t-on fait en tapant ces commandes? Que concluez vous à partir des résultats obtenus? Est-ce surprenant?